European Journal of Epidemiology (2005) 20: 563–564
DOI 10.1007/s10654-005-4410-4

EDITORIAL

Interaction: A word with two meanings creates confusion

Anders Ahlbom & Lars Alfredsson
Institute of Environmental Medicine and Stockholm Center for Public Health, Box 210 171 77, Stockholm, Sweden

Accepted in revised form 22 March 2005

Perhaps more than any other word in epidemiology, 'interaction' presents a challenge to clinical and epidemiological researchers. The problem stems from its applicability to describe two different phenomena. On the one hand, interaction refers to the biologic interaction of two or more causes of disease that together assert their influence on disease risk. On the other, interaction refers to statistical interaction which is the necessity for a product term in a linear model. In this editorial, we have two related goals: (1) To provide authors with a common terminology for discussing biological and statistical interaction; and (2) To emphasize statistical interaction in the additive model as the basis for assessing biological interaction.

Most causes of disease are dependent on the presence of other factors to assert their effects and this is biological interaction. Examples of how risk factors interact are ubiquitous in medical and public health sciences. One of the more striking examples is phenylketonuria (PKU), a metabolic disorder in which the combination of a genetic mutation and an environmental factor, in this case dietary exposure to a particular amino acid gives rise to mental retardation in children. Because exposure to both factors is necessary for PKU to occur, infants with the genetic defect are put on a life-long restricted diet and are able to prevent the development of the disease. Another example is human papilloma virus (HPV) and cervical cancer. HPV appears to be a necessary cause, since its presence is observed in $>99\%$ of cases, but not sufficient, since additional factors are required in order to ultimately lead to disease. These two examples from the medical literature provide a framework for understanding the concept of biologic interaction, where two risk factors are involved in the same sufficient cause of disease, or pathway towards disease.

Statistical interaction refers to a different concept altogether. It refers to the need to include an interaction, or product, term in a statistical model for the model to fit the data well. Statistical interaction terms may be included in a variety of models that express different relations between the involved variables. Models are often converted to linear models through suitable transformations. In a simple linear regression model, inclusion of an interaction term would imply that the relation between the independent and dependent variables is no longer additive. A logistic regression model on the other hand is implicitly exponential and thus multiplicative. It becomes additive only after a logarithmic transformation. As a consequence, the inclusion of an interaction term in the logistic regression model implies that the investigated relation is no longer multiplicative.

The confusion around the dual meaning of the term interaction has arisen in parallel with the widespread use of statistical modeling and software development for epidemiological research. With such tools different models can be fitted with great ease and interaction and other features can be assessed even without full consideration of the implications. Many of the models that are used for epidemiologic analysis are inherently multiplicative, including the probably most common of them all, logistic regression. As a consequence, the vast majority of analyses of epidemiologic data are based on a multiplicative model and hence, most analyses of statistical interaction in epidemiologic data are implicitly utilizing the multiplicative scale.

At the same time, the additive model underpins the methods for assessing biological interaction. Probably the best tool to comprehend this is the 'pie model' introduced to epidemiology by Rothman [1]. According to this model two component causes, or risk factors, are independent if no sufficient cause involves both of them; in other words, when two risk factors are independent there is no pathway to disease that requires the involvement of both risk factors. Conversely, two risk factors assert biological interaction if they are involved in the same sufficient cause, i.e., if there is at least one pathway towards disease in which both risk factors are required. Rothman has shown algebraically, based on disease rates connected to the pie model, that independent risk factors adhere to an additive model and that biological interaction results in departure from additivity of the disease rates [2]. Therefore, the empirical criterion to be employed when assessing biological interaction, is whether or not disease rates are additive. The additivity criterion stems from a basic, but general definition of biological interaction and is in no way an arbitrary choice. For an in depth discussion of this topic, see for example the recent text book by Rothman [2].

Consider the following example: A disease risk, per 100,000, is 1 for those who are unexposed to two risk factors A and B; it is 2 for those exposed to risk factor A but not to B, and it is 5 for those unexposed to A but exposed to B. One may now ask the question what the disease risk would be for those who are jointly exposed to both risk factors assuming that there is no biological interaction. Since independence implies that the disease risks are additive the risk for the jointly exposed would be: $2 + (5-1) = 6$. Therefore, if the risk for the jointly exposed was, say, 10, one would conclude that the two risk factors are not independent, or in other words that biological interaction is present. Whether or not statistical interaction is present or not is an entirely different issue and depends on the choice of model, or scale. With an additive model statistical interaction is present, since: $2 + (5-1) = 6 \neq 10$. However, there is no interaction on the multiplicative scale, because: $2 \times 5 = 10$.

Even though the arguments that lead up to the additivity criterion are straightforward and the criterion easy to apply, one often encounters other methods to analyze whether or not risk factors interact biologically. Particularly common is to test for multiplicativity, often implicitly and perhaps even without noting it. The reason is undoubtedly that the majority of the statistical models that are used by epidemiologists are constructed as exponential models and thus are inherently multiplicative. When statistical interaction is assessed in such models it is equivalent to assess multiplicatively, rather than additivity.

It is not the intention of this editorial to claim that multiplicative models have no role in relation to biological interaction. However, the role is not to assess presence or absence of biological interaction but rather to specify the nature of a particular biological interaction. For example, if cigarette smoking damages the cilia in the airways and thereby opens up for exposure from radon contaminated particles, one could hypothesize that as a result there would be multiplicative relation between cigarette smoking and radon exposure in relation to lung cancer risk. If such a multiplicative relation was observed it would not indicate absence of biological interaction. Rather, it would suggest that, indeed, there is biological interaction between radon exposure and cigarette smoking and in addition it would imply something about the nature of this particular biological interaction.

While the interaction term, in a logistic regression model, or some other implicitly multiplicative models, has no direct relevance for the issue of whether or not biologic interaction is present, biological interaction can still be assessed using such models. However, this requires that the model is defined in a special way and that the analysis is done adequately. Because logistic regression and Cox regression hazard models are so predominantly used in current epidemiology, an accompanying article in this issue explains how biological interaction analysis can be carried out from the results of these models with the help of common software [3]. In that article we also give reference to an Excel sheet that uses the output from a program for logistic regression or a Cox regression to perform biological interaction analysis.

The confusion regarding interaction would certainly have been minimized had different terms been used for the two different concepts biological and statistical interaction. Realizing how difficult it is to change terminology, a minimum requirement is that authors are crystal clear about the meaning of the term interaction when they use it and that they specify if they refer to biological interaction or statistical interaction. It may be helpful to explicitly write 'biological interaction' whenever that is the intended meaning. In analogy one would use 'statistical interaction' when referring to an interaction, or product, term in a statistical model. It will be the policy of the editors of EJE to strive for clarity in this matter and to support the use of unequivocal terminology.

Acknowledgement

We thank professor Hans-Olov Adami for valuable contributions during the preparation of this editorial.

References

1. Rothman KJ. Causes. 1976. Am J Epidemiol 1995; 141: 90–95; discussion 89.
2. Rothman KJ.. Epidemiology An introduction. Oxford University Press 2002: NewYork.
3. Andersson T, Alfredsson L, Kallberg H, Zdravkovic S, Ahlbom A. Eur J Epidemiol 2005; 20(7): 575–579.

Address for correspondence: A Ahlbom, Institute of Environmental Medicine and Stockholm Center for Public Health, Box 210, 171 77 Stockholm, Sweden
E-mail: anders.ahlbom@imm.ki.se